数据科学基础
(With Python)

描述统计学简介

统计学

- 统计学研究事物的数量方面,目的是从数量方面来认识事物.
  - 数量的多少
  - 数量间的关系
- 作为数据分析工具应用于自然科学,社会科学和工程技术的很多领域
- 分为描述统计学和推断统计学

描述统计学

- 描述统计学借助图表或概括性的数值来描述数据集的特征,使得数据集能以清晰可理解的形式得到展示
- 统计研究的起点是获得数据集,目标是探索这个数据集的数量特征.
- 如果获得了全体数据(总体),那么经过描述统计的处理,就达到研究目标了.

推断统计学

- 由于各种原因很难获得总体数据,只能得到样本.
- 推断统计:从样本包含的信息来推断总体的数量特征
- 描述统计和推断统计通常构成研究数据的前后两个阶段.
- 统计的中心任务是通过样本了解总体

基本概念(1)

- 数据集例

Name	Gender	Age	GPA
Zhao	male	18	3.0
Qian	male	20	2.7
Sun	male	20	2.9
Li	female	19	2.6

基本概念(2)

- 个体:统计分析的对象
  - 例:每个学生
- 总体:全体对象
  - 例:全体学生
  - 有限总体vs无限总体
- 特征:用来描述个体
  - 例:姓名,性别,年龄,平均绩点

基本概念(3)

- 变量:特征可视为变量
  - 对特征进行测量即得变量的值.
  - 变量的值
    - 随个体而变:例如不同学生的年龄
    - 随时间而变:例如一个学生的年龄
- 数据:对个体的一个或多个变量进行测量,所得的测量值.

基本概念(4)

- 总体的性质由全体个体的性质决定
  - 只需测量每个个体的特征即可了解总体
  - 无限总体或个体数目非常大的有限总体:不可能测量每个个体
  - 即使个体数目不是很大,但对个体特征的测量是破坏性的
- 样本:从总体中抽取的一部分个体

总体和样本的所指

- 术语总体和样本有两种用法
  - 研究对象
    - 总体:全体学生
    - 样本:部分学生
  - 对研究对象特征的测量值
    - 总体:全体学生的**GPA**
    - 样本:部分学生的**GPA**

如何得到数据

- 获得数据的两种方法
  - 观测:测量研究对象的特征
    - 不影响,不改变对象
    - 采集的数据类型受限制,因为观测者不能控制环境
  - 实验:对研究对象做实验,观测效果
    - 对照实验(A/B测试):对象分两组,置于不同环境,然后观测效果,作出选择.

抽样

- 如何确定样本成员及容量?
- 概率抽样:对象被选中的概率是已知的,且可能不一样
  - 随机抽样:个体被选中的概率都一样
    - 避免抽样偏见
    - 避免引入混淆因子:未被测量但影响被测量变量的变量
  - 不等概抽样:偏向某些对象
    - 避免系统性偏见

定性与定量变量

- 定性变量:类别型特征
  - 数据是类别值
  - **Name和Gender**
    - **Gender值:male和female**
- 定量变量:数量型特征
  - 数据是数值型的
    - 离散型变量:可能值有穷或可数,如**Age**
    - 连续型变量:实数区间,如**GPA**

用图表描述数据

- 描述数据集的数据分布
  - 测量值是什么
  - 每个值出现了多少次
- 展示数据集的数据分布
  - 统计表
  - 统计图

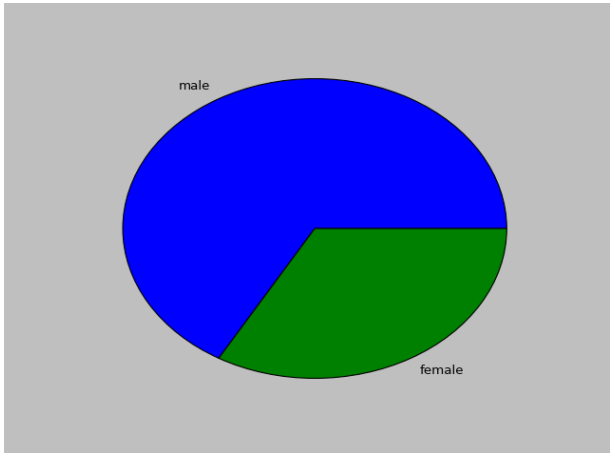
描述定性变量

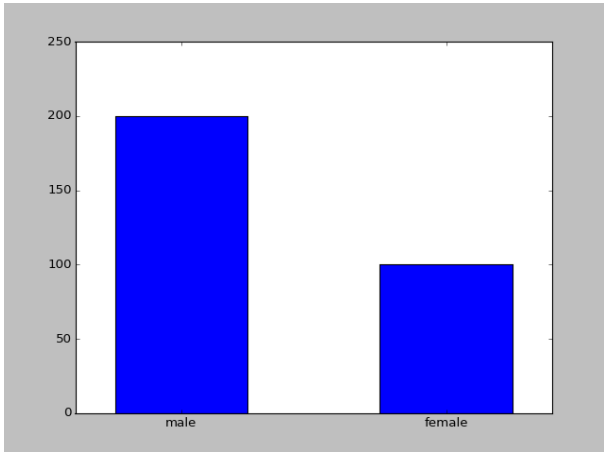
- 统计表:列出各值及出现频繁程度
- 数据值出现的频繁程度
  - 频数:数据集中出现了几次
  - 频率:出现次数/数据总数
    - [0,1]区间上的小数
    - 百分比

Gender	频数	频率	百分比
male	200	0.667	66.7%
female	100	0.333	33.3%
合计	300	1	100%

数据分布的可视化

Gender	频数	频率	百分比
male	200	0.667	66.7%
female	100	0.333	33.3%
合计	300	1	100%



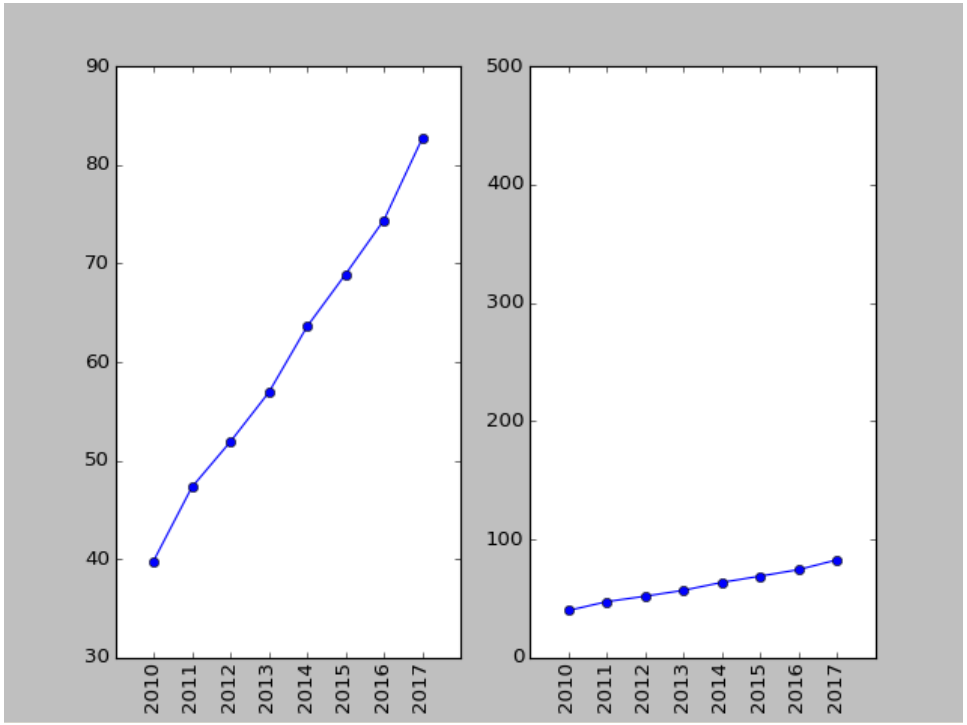


描述定量变量

- 分组考虑定量变量的值:类似定性变量
  - 用饼图或条形图来描述数据
  - 图中反映的是定量变量的值,而非某个类别的频数
  - 例如:针对年龄或性别分组,计算平均**GPA**
- 定量变量随时间变化时:等间隔地测量变量值,可得时间序列数据集
  - 最有效的展示方式是折线图
  - 从折线图发现模式或趋势并用来预测未来

描述定量变量

年份	2010	2011	2012	2013	2014	2015	2016	2017
GDP 万亿	39.8	47.3	51.9	56.9	63.6	68.9	74.4	82.7



频率直方图

- 频率直方图是条形图的一种
  - 将从最小值到最大值的区间划分成若干等宽子区间(桶)
    - 桶的划分要确保每个数据值落在唯一桶中
  - 在横轴上确定桶区间
  - 以桶中数据的频数或频率为高度绘制条形图
    - 频数直方图或频率直方图

桶的数目设定

- **经验规则:5到12之间**
  - **数据多则桶数多**
- **参考设定**

样本大小	25	50	100	200	500
桶的个数	6	7	8	9	10

例:频率直方图

- 数据集

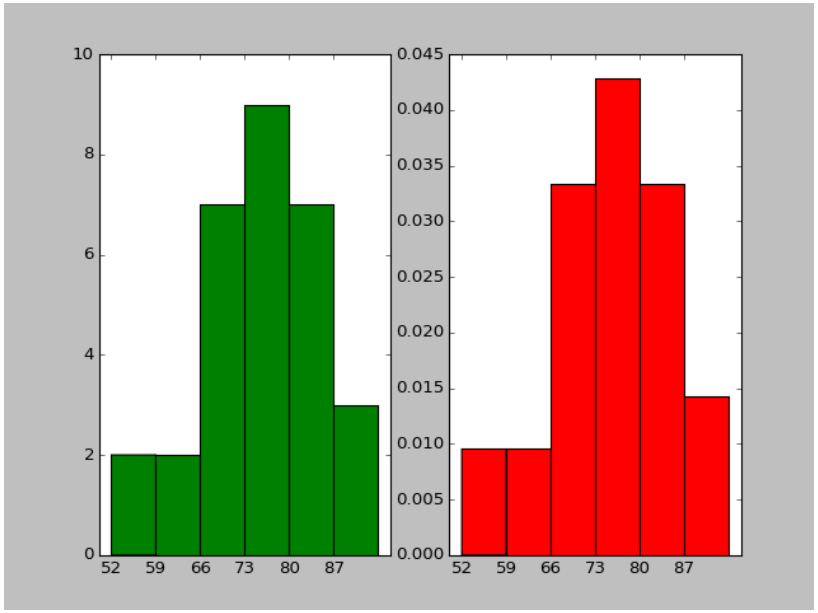
52 58 61 62 67 68 68 68 71 72 72 75 75 77 77
77 77 78 78 79 80 82 82 82 85 85 86 90 90 94

- 设划分6个桶
  - 桶宽度:(94-52)/6 = 7
  - 桶边界:52~59,59~66,...(左包括)
  - 计算频数频率

桶	频数	频率
52～59	2	2/30
59～66	2	2/30
66～73	7	7/30
73～80	9	9/30
80～87	7	7/30
87～94	3	3/30

例:频率直方图(续)

- 频数直方图和频率直方图



  - 频数/频率直方图类似,只是高度乘个因子
  - 右图实为频率分布直方图(高度是频率密度)

从直方图看数据分布

- 看直方图的位置和形状
- 前例中:
  - 两头小中间大
  - 单峰
  - 对称
  - 没有异常值

用数值描述数据集

- 用统计图描述数据集的数据分布不够精确,对基于样本推断总体是不够的
- 数据集的原始数据是描述个体特征的细节数据,而我们希望能有反映总体或样本特征的概括性数据
- 反映总体或样本数据分布的数值:对总体而言称为参数,对样本而言称为统计量.

集中性度量:均值

- 算术平均值

$$\frac{\sum_{i=1}^{n} x_i}{n}$$

  - 样本均值记为 $\bar{x}$,随样本而变.
  - 总体均值记为μ , 唯一.
- 均值是分布中心,观测值密集分布在均值两侧,呈现向心集中的趋势
- 均值对异常值敏感,可能带来误导.

集中性度量:中位数

- 中位数:将n个观测值从小到大排列,处于(n+1)/2位置的值就是中位数
  - n是奇数则有绝对中间位置
  - n是偶数则中间位置带小数.5
    - 这时中位数定义为中间位置两侧的两个观测值的算术平均值
- 一半值≤中位数,另一半值≥中位数.
- 对异常值不敏感
  - 例:[2, 5, 6, 9, 11, 45]的均值13,中位数7.5

集中性度量:中位数(续)

- 中位数对异常值不敏感
  - 例:[2, 5, 6, 9, 11, 45]的均值为13,中位数为7.5,更准确地描述了分布中心.
- numpy.median

```
>>> np.median([2,9,11,5,6])
6.0
>>> np.median([2,9,11,5,6,27])
7.5
```

集中性度量:众数

- 众数:数据集中出现次数最多的值
  - 一般用于较大的数据集
- 众数不受异常值的影响
- 有的数据集不止一个众数
  - 这通常说明数据可能来自两个总体
- **pandas的Series和DF有mode方法**

```
>>> s1 = pd.Series([1,2,2,2,3,3,4,4])
>>> s1.mode()
0   2
```

集中性与变异性

- 均值等统计量反映数据分布的集中趋势,但不能反映各数据的差异
  - 两个数据集可能具有相同的中心但分布差别很大,因为数据偏离中心的程度不同
- 变异或分散程度也是数据集重要特征
- 数据偏离中心不是数据好坏之分
  - 制造零件:不能偏离标准尺寸
  - 招聘面试:有偏差才好选择

变异性度量:极差

- 极差:最大值与最小值之差
  - 极差大,数据分散程度也大
  - 极差小,则数据值都很接近;为0则无差异
- 极差仅涉及两端,不能描述整个数据集
  - **[0,100] vs [0,50,50,50,50,100]**
  - 不适用大数据集

```
>>> np.ptp([5,7,1,2,4])
6
>>> np.ptp([[0,1],[2,3]],axis=1)
array([1, 1])
```

变异性度量:离差

- 离差: $x_i - \bar{x}$
  - 有正有负
- 平均离差?
  - 离差总和为0!
- 绝对离差:离差的绝对值
- 平均绝对离差:ok
  - 用平方来去掉负号更好

变异性度量:方差

- 方差:各数据值的离差平方的平均值
  - 总体方差

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

  - 样本方差

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

    - 为何样本方差的除数是**n–1?**
    - 用样本方差估计总体方差时,用**n–1**定义的方差能得到更好的估计

变异性度量:标准差

- 标准差:方差的正平方根
  - 总体标准差: σ
  - 样本标准差: s
- 方差的度量单位是原始数据度量单位的平方, 而标准差与原始数据度量单位相同
- **NumPy**

```
>>> np.var([5,7,1,2,4])
4.560000000000005
>>> np.std([5,7,1,2,4])
2.1354156504062622
```

标准差的意义

- 标准差(或方差)越大,说明数据的变化越大,分散度越大
  - 标准差为0,则所有数据值都相等,没有分散
- 切比雪夫定理:考虑n个数据值的集合.对任意$k \geq 1$,数据集的至少$(1-1/k^2)$落在$\mu \pm k\sigma$范围之内.

k	$1-1/k^2$	含义
1	0	至少0个值落在$\mu \pm \sigma$内
2	3/4	至少3/4的值落在$\mu \pm 2\sigma$内
3	8/9	至少8/9的值落在$\mu \pm 3\sigma$内

相对排位:标准分

- 一个观测值的绝对数值并非总是最重要的,很多时候我们更关心该值的相对排位.
- 标准分(z分数):以标准差s为单位来度量一个值x与均值 $\bar{x}$ 的差

$$z = \frac{x - \bar{x}}{s}$$

- 假设班级考试均分为65,标准差为5,则考75分的同学的标准分为(75-65)/5 = 2,意为该生位于超过平均成绩2个标准差的位置

标准分与异常值

- 根据切比雪夫定理:
  - 至少75%的观测值落在均值的2个标准差范围之内,即这些观测值的标准分介于-2到2之间
  - 至少89%的观测值落在均值的3个标准差范围之内,即这些观测值的标准分介于-3到3之间
  - 通常标准分绝对值大于3的观测值可以认为是异常值

例:标准分

- 标准分可以更客观准确地分析数据

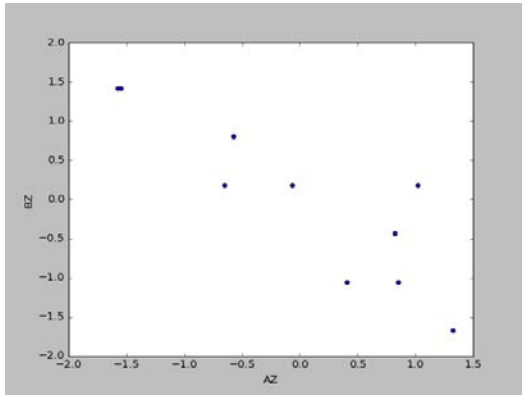
	语文	数学	外语	政治	物理	化学	总分
甲分数	70	57	45	80	60	70	380
乙分数	90	51	40	85	72	62	400
平均分	70	55	42	80	36	50	
标准差	8	4	5	10	12	4	
甲标准分	0	0.5	0.6	0	2	5	8.1
乙标准分	2.5	-1	-0.4	0.5	3	3	7.6

例:不同尺度的标准化

- 标准分将不同尺度的数据统一尺度

A	10	102	113	42	63	96	9	95	80	45
B	0.8	0.6	0.3	0.6	0.6	0.4	0.8	0.5	0.4	0.7
AZ	-1.5	1.0	1.3	-0.7	-0.1	0.9	-1.6	0.8	0.4	-0.6
BZ	1.4	0.2	-1.7	0.2	0.2	-1.1	1.4	-0.4	-1.1	0.8

```
import pandas as pd
from sklearn import preprocessing
a = [10,102,113,42,63,96,9,95,80,45]
b = [0.8,0.6,0.3,0.6,0.6,0.4,0.8,0.5,0.4,0.7]
df = pd.DataFrame({'A':a,'B':b})
dfZ = pd.DataFrame(preprocessing.scale(df),
                   columns=['AZ','BZ'])
dfZ.plot(kind='scatter',x='AZ',y='BZ')
```



相对排位:百分位数

- 设数据集的n个值从小到大排列
- 第p个百分位数:是这样一个数值,它比p%的测量值大,比其余(100-p)%的测量值小
- 常用的几个百分位数:
  - 第25个百分位数:下四分位数(或第一四分位数),记作Q_1
  - 第50个百分位数:即中位数
  - 第75个百分位数:上四分位数(或第三四分位数),记作Q_3

四分位

- **Q_1,中位数,Q_3将数据集等分为四个子集**
- **Q_1和Q_3是数据集的中间50%个数据的下界和上界**
- **四分位距IQR:$Q_3 - Q_1$**

四分位数的计算

- 小数据集可能不好四等分(如n=10),或者能等分(如n=12)但有多个值满足四分位数定义
- 计算规则:Q_1是位于0.25(n+1)处的值,Q_3是位于0.75(n+1)处的值
  - 如果位置不是整数,则利用与之相邻的两个数据值来计算Q_1和Q_3

例:四分位数

数据集[4, 8, 9, 11, 11, 13, 16, 18, 20, 25]

Q_1=0.25×(10+1)=2.75

取与此位置相邻的8和9间距的0.75处:

Q_1=8+0.75×(9–8)=8.75

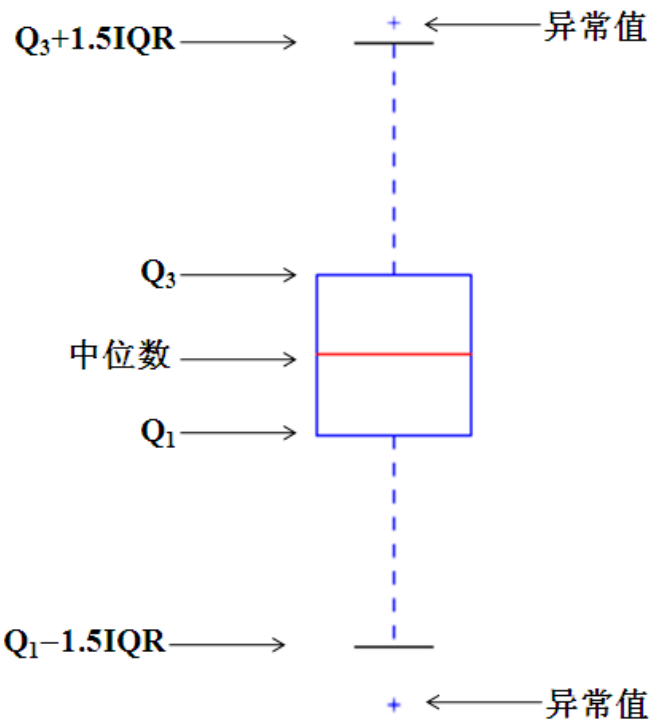
Q_3=0.75×(10+1)=8.25

取与此位置相邻的18和20间距的0.25处:

Q_3=18+0.25(20–18)=18.5

五数概括法

- 一个数据集可以用最小值,Q_1,中位数,Q_3和最大值来快速概括其数据分布情况
  - 用箱线图可视化呈现
  - 中位数居中:对称分布
  - 中位数近Q_1/Q_3:偏斜



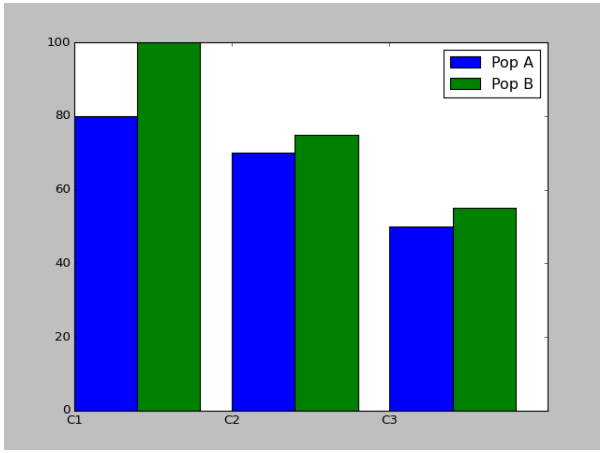
双变量数据

- 经常研究多个变量的相互关系
  - 家庭消费支出与家庭人口的关系
  - 房地产售价与房间数的关系
- 双变量数据:对个体同时测量两个变量.
  - 单独研究每个变量
  - 研究两个变量之间的关系

用统计图描述双变量数据

- 至少一个定性变量:适合用饼图,条形图等
  - 例如:用定性变量作为横轴,用另一个变量(定性或定量)作为纵轴, 绘制条形图
- 定性变量和定量变量:数据来自两个总体,可以用并列饼图,并列条形图,堆叠条形图描述.

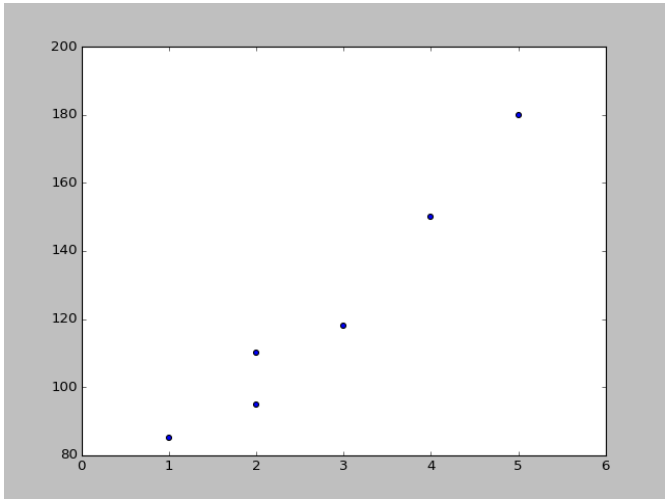
	C1	C2	C3
Pop A	80	70	50
Pop B	100	72	51



用统计图描述双变量数据

- 两个定量变量:散点图
  - 变量x作横轴,变量y作纵轴
  - 每一对对应值用点(x,y)表示
- 从散点图可观察两个变量之间的关系

x	2	2	3	4	1	5
y	95	110	118	150	85	180



用数值描述双变量数据

- 线性相关:散点图中所有点围绕一条直线分布
  - 一个变量每增减一个单位,另一变量按固定的增减量变化
- 均值和方差等不能描述两个变量之间的关系

协方差和相关系数

- 协方差

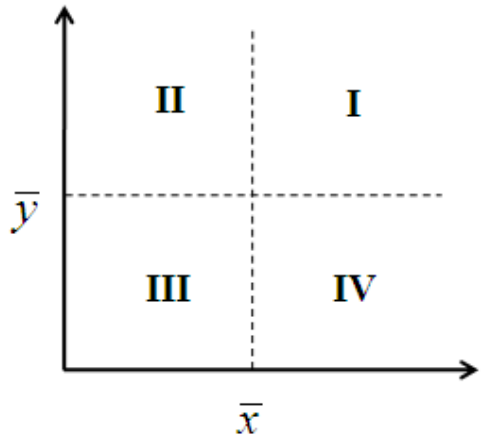
$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

  - 方差度量单个变量偏离均值的程度
  - 协方差度量两个变量联合地偏离其平均值的程度
- 相关系数

$$r = \frac{s_{xy}}{s_x s_y}$$

协方差和相关系数的直观意义

- 如果数据主要分布在第I、III象限, 则s_{xy}和r都大于0，这时存在从左下到右上的直线模式
- 如果数据主要分布在第II、IV象限,则s_{xy}和r都小于0，...
- 如果数据点在四个象限到处分布，这时s_{xy}和r都接近于0



例:计算协方差和相关系数

- **NumPy提供cov和corrcoef函数**

```
>>> xs=[2,2,3,4,1,5]
>>> ys=[95,110,118,150,85,180]
>>> np.cov(xs,ys)
array([[   2.16666667,   51.6       ],
       [  51.6       , 1280.        ]])
>>> np.corrcoef(xs,ys)
array([[1.        , 0.97982534],
       [0.97982534, 1.        ]])
```

  - 主对角线分别是x和y的方差(相关系数)

讨论:相关≠因果

- 因果关系:一个事件发生导致另一事件发生,原因在前结果在后
  - 例如:家庭收入与存款
- **x与y相关:并不意味着x和y之间存在因果关系**
  - 例如:医院数量与盗车数量线性相关
  - 这两个数量是由人口数导致的

讨论:没有线性相关≠不相关

- 如果相关系数为0, 则两个变量之间不存在线性相关——仅此而已
  - 根据x值偏离 $\bar{x}$ 的信息了解y值偏离 $\bar{y}$ 的情况
- 不存在线性相关, 并不意味着没有其他类型的相关

x	–2	–1	0	1	2
y	2	1	0	1	2

  - 关系:y是x的绝对值

讨论:Simpson悖论

- 双方在比较两组数据时,在分组比较中各组都占优势的一方,却在不分组的总体比较中处于劣势.（权重问题）

学院	女申请	女录取	女录取率	男申请	男录取	男录取率
商学院	100	49	49%	20	15	75%
法学院	20	1	5%	100	10	10%
总计	120	50	42%	120	25	21%

End